Ameriflux Data Curation Portal Status and Learnings to Date

Berkeley Water Center Microsoft TCI

July 2006

Outline

- Ameriflux Data Overview (reminder)
- Tales from our June download
- Portal overview – the vision thing
- Next steps

Ameriflux Overview

- 149 Sites across the Americas
- Each site reports a minimum of 22 common measurements.
- Communal science – each principle investigator acts independently to prepare and publish data.
- Data published to and archived at Oak Ridge.
- Total data reported to date on the order of 110M half-hourly measurements.

- http://public.ornl.gov/ameriflux/

Why chose Ameriflux ?

- Lots of different measurement types
- Time series analysis very important
- Bridge to other data sets (e.g. MODIS)
- Not self contained data - scientists add own derived measurements
- Good practice for sensor data to come
- Expect to be able to repurpose the Ameriflux portal to other water or environmental data sets

Tales from our June download

Disclaimer

The next several slides talk about our recent experience downloading data from the current Ameriflux web site.

This discussion is NOT a criticism of the current web site.

We've looked at others.
We talked to other scientists in different disciplines.
We've talked to other users.

The current Ameriflux site and the collaboration cooperation is much better than most.

June Data Download

- Using a simple bot, we downloaded all data available from http://cdiac.esd.ornl.gov/programs/ameriflux/data_system/aamer.html
- Goals for the download:
  - "Complete" data set of "known" provenance
  - Basis for technology evaluation and scaling studies
  - Useful for research – continue getting feedback

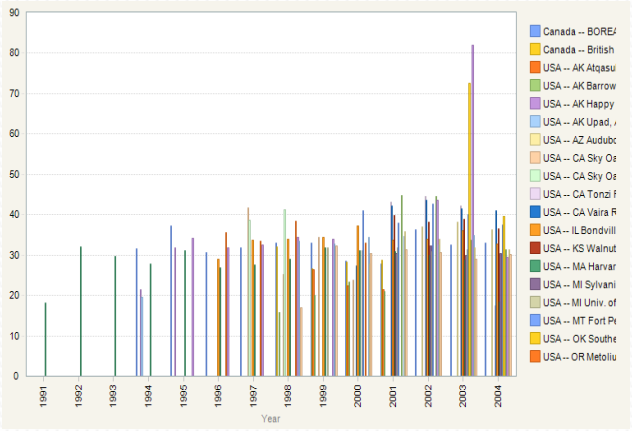
June ORNL database

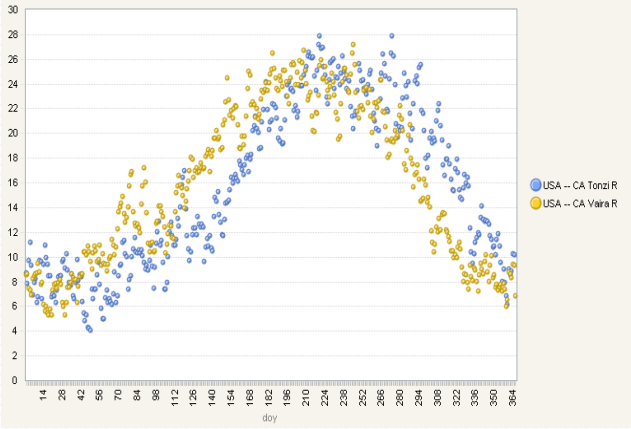
- 61 sites, 32 researchers, 15 years, 110M (non-gap) values.
  - 627 unique column names (data types) because of uncontrolled qualifiers added to 40+ primary data types. SWC2_10 or WS_cup2_2_5
  - Roughly 1 in 5 files failed N database ingest due to extra commas or line feed in column headers.
  - 17 missing sites – sites with submitted data not on web page. We know some are available off other ftp sites.
  - About to run basic Q/A checks

Curation starts with Q/A

- Getting data clean is a never ending task
  - Erroneous units
  - Calibration corrections (eg bird poop)
  - Time shift errors
  - Cut and paste human errors
- Some are easy checks
- Some are best spotted when compared with other data

Implies a need for data versioning and constant feedback between researchers and the portal

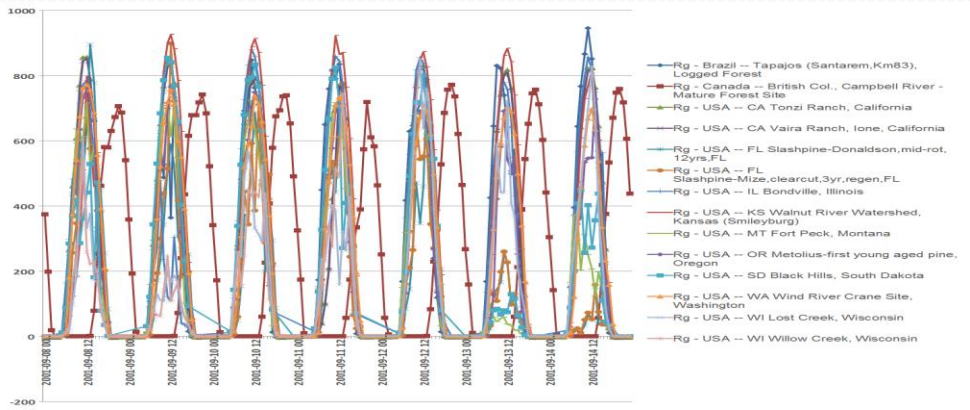


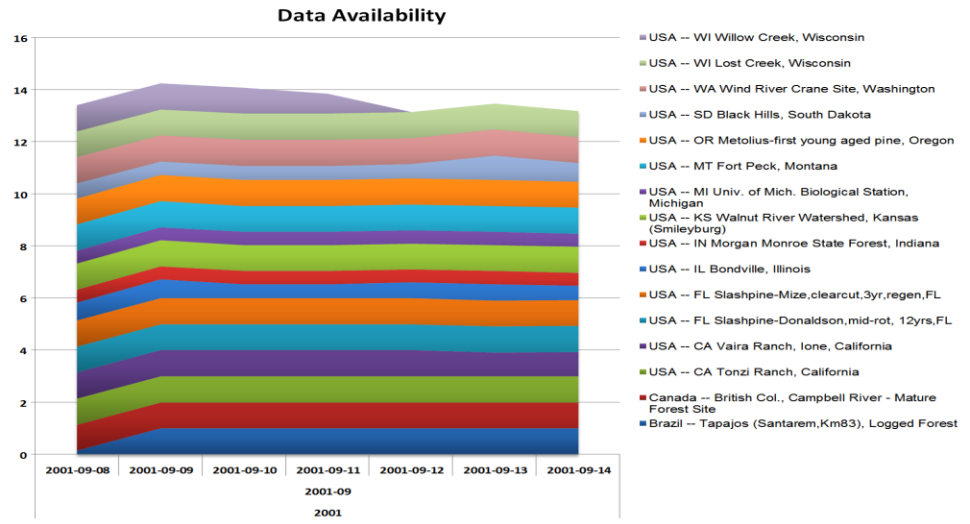


Analyzing the whole dataset

- Query large quantities of data and reduce to smaller subset for careful rendering
- Query and render large quantities of data for browsing "what if" viewing

Implies that the database back end must be matched with the "right" viewers





Portal Overivew
The Vision Thing

Portal Architecture

- Portal supports specific work flows
  - Data upload, download, and curation
  - Large data set analysis and browsing
- Built on a collection of distributed components
  - Extensible with csv data interchange
  - Leverage commercial software technologies

Portal Workflows

- Data download
  - Similar to today - select sites, datum types, time ranges
  - CSV files(s) with controlled header text
- Data upload
  - Ingest into a staging database and cube
  - Basic Q/A including browse viewing before data published to archive
- Data browsing
  - Multi-site and multi-measurement type plotting
  - Can download contributing or plotted results
- Private data analysis
  - Personal or shared data subsets
  - Allows creation of new measurement types (e.g. model results)
- Collaboration membership and site management
  - Supports data change notifications, data reporting, private data analysis resource tracking

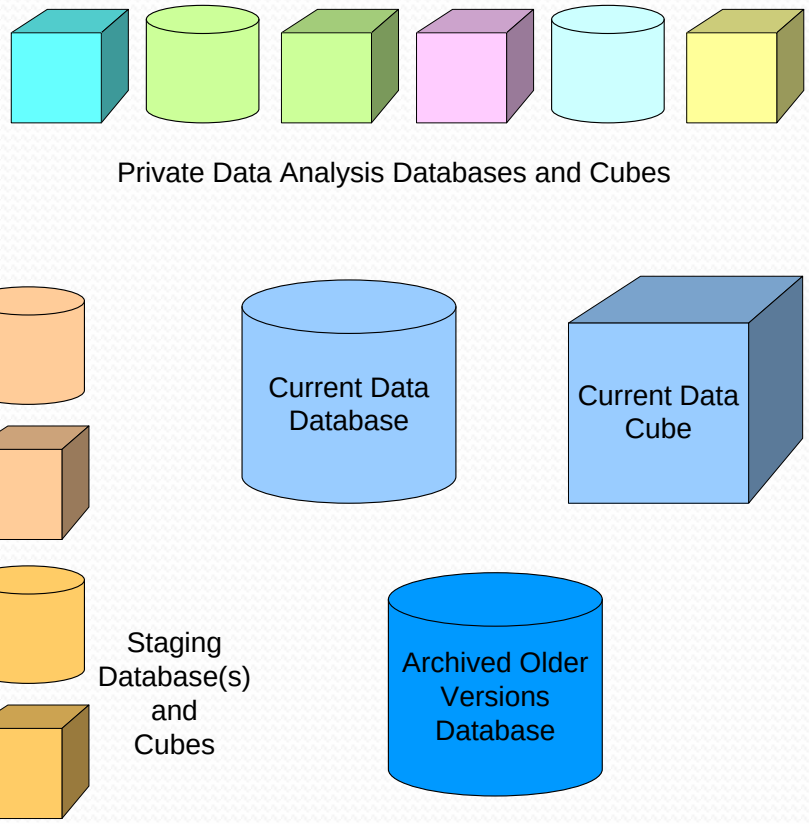
Portal Browse and Desktop Drill Graphs

- At selected sites, simple calculation over a time period plotted over a time range.
  - Daily sum of LE at Tonzi and Vaira vs time during 2000 to 2004
- At selected sites, scatter plot simple calculation over a time period over a time range
  - Maximum hourly temperature at Tonzi vs Metolius during 2001 and 2004
- Probability density functions
  - Distribution of half-hourly SWC measurements at Tonzi and Vaira
- Diurnal variations
  - ??? We need some advice here
- Color coded tile plot of a simple calculation at a single site over a time period
- Correlation of a simple calculation at a site over a range of lag times over a time period
- "Folded time" correlations and statistics
  - After more than 2mm rain…

We need your feedback on these – they form the test cases for our technology studies

Portal Deployment

- Behind the portal are a collection of databases and data cubes
- Distribution for ease of use
  - Only see the data of interest
  - Private data remains stable
- Distribution for scaling
  - Smaller queries on smaller databases take less resources
  - Larger databases and cubes can be replicated across machines
- Batch job like infrastructure for managing very log running queries



Next Steps

Next steps

- Support Gretchen, Dennis, Bev, and ?
  - We need your feedback on tools, usability, etc
  - We need a list of graphs/questions/queries you want to be able to do
- Build portal infrastructure
  - Big data browse graphs
  - Basic upload/download CSV interchange
  - Walmart vs Skyscraper, tables or cubes tradeoffs
  - “My Cube” generation infrastructure

Microsoft®

Your potential. Our passion.™